DOCUMENT RESUME

ED 051 302 TM 000 624

AUTHOR Sabers, Darrell L.; White, Gordon W.
TITLE The Effect of Differential Weighting of Individual Item Responses on the Predictive Validity and Reliability of an Aptitude Test.
PUB DATE 71
NOTE 11p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Computer Oriented Programs, Item Analysis, Measurement Techniques, *Multiple Choice Tests, Predictive Measurement, *Predictive Validity, *Scoring, Scoring Formulas, *Test Reliability, Tests, Test Validity, *Weighted Scores

ABSTRACT

A procedure for scoring multiple-choice tests by assigning different weights to every option of a test item is investigated. The weighting method used was based on that proposed by Davis, which involves taking the upper and lower 27% of a sample, according to some criterion measure, and using the percentages of these groups marking an item option to obtain the weight for that option. These percentages were then used to enter a weighting table to derive the appropriate option weight. Weights assigned to one item need not be similar to those of another item; an incorrect response to a difficult question may carry more weight than the correct response to an easier question. Weights for scoring the Iowa Algebra Aptitude Test were determined by computer by the use of achievement tests and the IAAT itself given to two groups of ninth grade algebra and two groups of ninth grade modern mathematics students. Correlations between the pairs of weights were used as measure of the reliability of the choice weights. The data suggests that more than 1,000 examinees would be required to provide reliable scoring weights for the distracters in this test. The cross-validation of the weights indicates a limited increase in both predictive validity and reliability. It is suggested that the main utility of the technique may be to increase reliability where greater reliability of measurement is needed. (DG)

ED051302

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

THE EFFECT OF DIFFERENTIAL WEIGHTING OF INDIVIDUAL ITEM RESPONSES ON THE PREDICTIVE VALIDITY AND RELIABILITY OF AN APTITUDE TEST

Darrell L. Sabers and Gordon W. White
Bureau of Educational Research and Service
The University of Iowa

With the application of the computer to the process of scoring objective tests and with the existing possibility of computer scoring allowing the development of new item types, it is perhaps time to re-evaluate present methods of scoring multiple-choice tests.

Typically, multiple-choice items are scored with unit weights given to correct item responses and with constant weights (either zero or, if a correction for chance success is employed, the weight $-1/(k-1)$, where k equals the number of options for each item) assigned to incorrect item responses. Other item scoring procedures have been presented in the literature, but their application to practical testing situations has been slow. One such procedure, employing choice weights, though a tedious task in the hand scoring age, appears to be very practical in the machine scoring age.

Choice weight scoring refers to the procedure whereby different weights may be assigned to all options of an item. For an item with four options, for example, the correct response may be assigned a weight of +3 and the incorrect options may be assigned weights of 0, -1, and -3. Another item in the same test may have a weight of +1 for the correct answer and weights of 0, 0, and -2 for the foils. That is, the weights assigned to any one item need not be similar to those assigned another item. It is obvious that a completely non-discriminating item should not be included in the test; however, non-discriminating options are often included, and zero weight is assigned to these options.

Although this type of scoring has seldom been employed with objective tests, it is by no means a unique idea in educational institutions. The teacher who employs essay tests in evaluation has long given different amounts of credit for answers differing in degree of correctness. The Strong Vocational Interest Blank employs weighted scoring, though it is an inventory type check list rather than an achievement examination. The method employed in this study is simply an extension of a similar scoring procedure to objective tests.

Scoring formulas that assess partial knowledge have been proposed and used by Coombs, Milholland, and Womer (1956) and by Dressel and Schmid (1953). In these studies it was concluded that partial knowledge does exist and that by employing proper scoring techniques the reliability of multiple-choice tests may be increased. Ferris (1967) concluded that differential weighting of correct responses deserves further study, even after his study failed to provide

TM 000 624

evidence that it increases validity. Although the primary purpose of the method of scoring employed in the present study is to differentiate among the examinees who are unable to identify the correct choice, it also provides for differential weighting among the correct choices.

Moore (1956) and Blood (1951) assigned a priori weights and raised test reliability without changing validity. Davis and Fifer (1959) have shown that choice weight scoring can increase the reliability of a test with no decrease in concurrent validity. The greater reliability of the weighted scores is attributed to the addition of variance resulting from the examinees' selecting among incorrect choices. Davis and Fifer stated that the process of selecting among incorrect options apparently measures the same mental functions as the selection of the correct choice from among all the options in the item. However, if items measuring the same function are added to a test to increase its reliability, one expects an accompanying increase in validity. Unless the increments in reliability due to choice weight scoring are accompanied by an increase in criterion-related validity, it may be assumed that what is being added by using choice weights is non-relevant variance.

Though the above-mentioned studies all show an increase in reliability resulting from choice weight scoring, none has demonstrated an increase in validity. This may have resulted because an increase in reliability was the primary objective in most cases. In the present study, an increase in validity was the primary objective, and reliability was considered secondary.

The present study was designed to assess the effect of choice weight scoring on predictive validity by determining if weights derived for one group can be successfully cross validated. It is to be expected that if the weights for the items of a test are to be useful, a cross validation should indicate that these weights have predictive merit with a similar group. In this study, a set of weights determined for one group was used with three other groups.

Each of the four groups in this experiment contained 370 students enrolled in junior high schools in Iowa. Since some schools classify their ninth grade offerings as modern mathematics and others as traditional algebra, two groups (M1 and M2) were chosen from the former category and the other two groups (A1 and A2) were selected from the latter category. The students in all four groups were administered the Iowa Algebra Aptitude Test while in the eighth grade. As criterion measures, forty-item multiple-choice achievement tests were administered after the students had completed one semester in ninth grade mathematics. Different, but similar, achievement tests[1] were used for the modern mathematics and algebra groups. The achievement tests were scored number correct.

Many methods of scoring multiple-choice tests may be employed to assess partial knowledge. The method used in this study is essentially that proposed by Davis(1959), which is a simplification of that proposed by Flanagan(1935). The upper and lower 27% of a sample are chosen according to some criterion measure. Then the percentages of these groups marking an item option are used as arguments to obtain the weight for that option from the table prepared by Davis (1966, Table VII). For example, if fifty per cent of the upper 27% chose

---

[1]These achievement tests contained items which are now included in the Kepner Mid-Year Algebra Achievement Test, currently being developed by the Bureau of Educational Research and Service.

option "a" for item 2, and if twenty-four percent of the upper 27% chose this option, the percentages 50 and 24 are used to enter the Davis Table. For this example the choice weight for option "a" would be +3. In scoring the test, a student marking this option gets a score of 3 points for item 2.

Within each of the four samples, the upper and lower 27% were chosen first on the basis of scores on the criterion (achievement) test. The procedure was then repeated with the upper and lower 27% chosen on the basis of the aptitude test scores (where the score on the test was simply the number of items answered correctly). Thus, eight different sets of weights were obtained for the aptitude test: four sets were based on groups selected on the basis of aptitude test scores. Thus, there were nine scorings of the IAAT: once using the formula score equals number right, and once for each of the eight determinations of the choice weights.

The actual determination of the choice weights was done entirely by computer. The answer sheets containing responses to the aptitude test, as well as the scores on the criterion measure, were read (not scored) by an IBM 1230 optical scanner. This information was then transferred to an IBM 534 keypunch attached to the IBM 1230 and punched on standard IBM cards. The cards were then fed into an IBM 7044 computer. For the purpose of this experiment, an existing item analysis program was adapted to choose the upper and lower 27% of each sample on the bases previously described and to compute the percentage in each group that chose each option. The optimum weight for each option was then determined from the Davis Table which had been read into the computer. The program was cycled twice to select the upper and lower groups from scores on the achievement test and the IAAT, respectively.

Although samples of 370 examinees were employed in this study, the method of scoring tests by computer can utilize any reasonable number. Sample size was determined in order to assess the feasibility of employing weights derived from a sample of this size since practical considerations (Davis, 1959) often suggest using multiples of 370 examinees.

It should be noted that the present method of determining weights differs somewhat from that employed by Davis (1959). This study used as weights integers ranging from -9 to +9 as read from the Davis Table. Davis (1959) used a transformation which provided as scoring weights integers ranging from -3 to +3. Since a high speed computer and not the IBM 1230 was used to score these tests, there was no need to make such a transformation. In addition, Davis employed upper and lower groups based only on the total scores of the test for which choice weights were to be determined. On the other hand, in this study an outside criterion was used to select the groups for four of the sets of weights in order to study the increase in predictive validity.

Obvious objections arise to assigning choice weights entirely on the basis of empirical increments to validity or reliability. In the procedure employed in this study, it is possible to get positive credit for a wrong answer--in fact, it is possible to get more credit for a wrong answer than for a correct answer to that item or to another item. It may have been partly for this reason that Davis and Fifer used "modified" weights (they contend, rightly so, that the moderate reliability of the empirical weights is reason to modify

weights by a priori reasoning). However, it is the contention of the present writers that a subject may exhibit more ability by selecting option 'B' for item 32 (an incorrect choice to a difficult item) than by selecting option 'B' for item 1 (a correct choice to an easy item). In this case, it is perhaps desirable to have a positive weight for 32-B and a zero weight for 1-B. Following this reasoning, no modification was made of the empirical weights.

Reliability of estimated weights

In order for choice weight scoring to be worthwhile, it is imperative that the weights be reliably determined. Independently estimated scoring weights were obtained for each of the 320 options of the aptitude test. When the upper and lower 27% were selected on the basis of the achievement scores, a set of weights was obtained for each of the two modern math groups. Thus, each option had two weights assigned, one for each modern math group. The correlation between these pairs of weights is a logical measure of the reliability of the choice weights. Since weights for the correct responses are distributed around a positive mean and weights for distracters are distributed around a negative mean, correlations were computed separately for the correct choices and for the distracters.

The correlation between the weights assigned to the correct responses for the two modern math groups was .85; that for the distracters was .39. This procedure was repeated for the algebra groups; the coefficients were .87 and .35, respectively. The higher stability of the correct responses is to be expected because of the larger number of examinees who choose correct responses as compared to any incorrect answer, and because guessing may play more of a part in the choice of distracters.

Table 1 presents the correlations between the various pairs of weights assigned to the options when the upper and lower 27% of each sample was chosen on the basis of the aptitude scores. The correlations between pairs of weights assigned to the eighty correct choices are reported above the diagonal in Table 1, and those for the two hundred and forty distracters are reported below the diagonal. The correlation between the weights assigned to the correct responses for the two modern math groups was .87; for the distracters it was .47. The correlations for the algebra groups were .84 and .50, respectively.

Since grouping into the modern math and the algebra categories was done on the basis of the ninth grade courses, it is possible that the four samples of students were quite similar in mathematics background at the time of the administration of the IAAT in grade eight. Therefore, correlations between all pairs of sets of choice weights[2] are reported in Table 1. The coefficients for the correct responses ranged from .84 to .91, those for the distracters ranged from .36 to .54.

---

[2]When upper and lower 27% were chosen on the basis of IAAT scores.

To examine the effect of sample size on the reliability of the weights, additional sets of weights were determined for the combined algebra groups. The correlation between these weights and weights determined for the combined modern math groups (each set based on 740 subjects) was .95 for the correct responses and .62 for the distracters. Thses data suggest that well over 1000 examinees would be required to provide reliable scoring weights for the distracters in this test. However, a sample of lower ability students (or use of a more difficult test) might provide a more accurate estimation of weights.

Validity of choice weight scoring

Table 2 presents data showing the effect on predictive validity of employing weights obtained when the upper and lower 27% were chosen on the basis of achievement test scores. For each group, the first row presents validity coefficients (and increments) with the achievement test as criterion. The second row presents Spearman-Brown (odd-even stepped up) estimates of the reliability of the aptitude test. Each column in Table 2 presents data based on a different scoring of the test. For example, the validity coefficient for group M1 (modern math group one) was .767 for the aptitude test scored number right. When weights derived from the other modern math group were used for group M1, the increase in the validity coefficient (over .767) was .004. When weights from group M1 were used to score the test for group M2, there was no change in the validity coefficient. For the algebra groups, the cross validation increments were .023 and .025.

It may be noted that when weights from groups M1 and M2 were used to score the aptitude tests for groups A1 or A2, the increments ranged from .020 to .036. However, when the weights derived from the algebra groups were used with the modern math groups, the increments ranged from -.001 to .012. One may ask why the results of the cross validations were so inconsistent. The writers have no explanation.

The increments in reliability were smaller than those in predictive validity. This is to be expected because of the relatively high reliability of the aptitude test. As reported in Table 2, the cross validation indicates that little increase in reliability resulted from the use of the weights. Neither in modern math (.004 and .015) nor in algebra (-.006 and .013) were increments great enough to have any practical significance.

Table 3 is similar to Table 2 except for the basis by which the weights were determined. Table 3 presents the results from scoring the aptitude test when weights were obtained by selecting the upper and lower 27% on the basis of aptitude test scores. A comparison of the increments in validity and reliability reported in Table 3 with those reported in Table 2 shows no marked superiority for either method of obtaining weights.

Table 4 presents the means and standard deviations of the weights assigned to each group. Table 5 presents data which indicate that the groups were not as comparable as might have been the case if subjects rather than schools had been the unit of random assignment. These data indicate that Group M2 had a higher level of ability and was a less variable group than the others. This

may be a reason why the cross-validation of weights in modern mathematics showed very little evidence of usefulness. The failure with the modern math group indicates that one cannot generalize about the usefulness of weights for a group other than the one for which the weights are derived.

It is possible that mathematics is the poorest possible area in which to defend weighted scoring, since the correctness of an answer is more clear-cut in this area than in most others. Two reasons besides availability of data can be offered for the writers' choice of this area. The test used in this study, the IAAT, includes 34 incomplete sequences in which the student is to determine the next term. While usually there is agreement among experts as to the best answer to such an item, a student may find a valid but unanticipated rule for a sequence which produces a response other than the keyed answer. Also, it was felt that if choice weight scoring were proven effective in the area of mathematics it should be even more likely to be fruitful in most other areas.

One additional point seems worthy of mention. In scoring with choice weights, one is not forced to use the computer. Davis (1959) suggested how the IBM 1230 optical scanner can be conveniently employed to use choice weights. If the weights employed range from, say, -2 to +2, one could handscore the answer sheets by using four separate stencils. This would require four scorings of each answer sheet, but many instruments in the areas of personality and interest assessment require multiple scorings. Where the computer is the real time saver is in the determination of the choice weights.

Conclusions

The results of this study seem to support previous research. Even though the chief purpose of the study was to increase test validity, this was accomplished only to a limited degree with the algebra groups. A tentative hypothesis could be that choice weight scoring serves only to add non-relevant variance, and thus is measuring a mental function different from that of selecting the correct response from among all options. Though this hypothesis has not been subscribed to by previous writers, it may be implied from the results of previous studies.

Most standardized achievement tests have a relatively high degree of reliability. The use of choice weight scoring might prove to be more fruitful in other areas of assessment where greater reliability of measurement is needed. It would be interesting to experiment with choice weight scoring for tests measuring analysis, synthesis, and evaluation, especially when these tests employ "best answer" items rather than items which have only one correct response.

REFERENCES

1. Blood, D. F., "The effect of the Use of the SRA Self-Scorer in the Measurement of Spelling Ability." Unpublished Doctoral Thesis, University of Iowa, 1951.

2. Coombs, C. H., Milholland, J. E., and Womer, F. B. "The Assessment of Partial Knowledge." Educational and Psychological Measurement, XVI (1956).

3. Davis, F. B., Analyse Des Items. Louvain, Belgium: Nauwelaerts, 1966.

4. Davis, F. B., "Estimation and Use of Scoring Weights for Each Choice in Multiple-Choice Test Items," Educational and Psychological Measurement, XIX (1959), 291-298.

5. Davis, F. B. and Fifer, G. "The Effect of Test Reliability and Validity of Scoring Aptitude and Achievement Tests with Weights for Every Choice." Educational and Psychological Measurement, XIX (1959), 159-170.

6. Dressel, P. L. and Schmid, J. "Some Modifications of the Multiple-Choice Item." Educational and Psychological Measurement, XIII (1953), 574-595.

7. Ferris, M. J. "Validity as a Function of Empirical Scaling of Test Items by a Logistic Model." Educational and Psychological Measurement, XXVII (1967), 829-835.

8. Flanagan, J. C. Factor Analysis in the Study of Personality. Stanford: Stanford University Press, 1935.

9. Moore, R. "A Comparison of Selected Modifications of a Multiple Choice Examination." Unpubliched Doctoral Thesis, University of Iowa, 1956.

TABLE 1

Correlations between sets of choice scoring weights* based on 4 independent samples of 370 examinees each (coefficients above the diagonal are for the correct responses (n=80), below diagonals are for distracters (n=240).

	GROUP			
GROUP	M1	M2	A1	A2
M1		.87	.87	.85
M2	.47		.84	.91
A1	.54	.36		.84
A2	.52	.52	.50	

*Upper and lower 27% chosen on the basis of aptitude scores.

TABLE 2

Validity and Reliability coefficients for the Aptitude test scored number right and increments resulting from choice weight scoring.*

Group	Type of Coefficient	Correlation S=R	Increment for weights derived from Group M1	M2	A1	A2
M1	Validity	.767		.004	.012	.008
	Reliability	.891		.004	.015	.003
M2	Validity	.713	.000		-.001	.002
	Reliability	.871	.015		.016	.007
A1	Validity	.745	.020	.032		.023
	Reliability	.875	.011	.001		-.006
A2	Validity	.674	.032	.036	.025	
	Reliability	.883	.009	.005	.013	

*Upper and Lower 27% chosen on basis of achievement test scores S=R.

TABLE 3

Validity and Reliability coefficients for the Aptitude test scored number right and increments resulting from choice weight scoring.*

Group	Type of Coefficient	Correlation S=R	Increment for weights derived from Group M1	M2	A1	A2
M1	Validity	.767		+.012	.006	.011
	Reliability	.891		.007	.018	.010
M2	Validity	.713	-.007		-.010	.002
	Reliability	.871	.017		.015	.015
A1	Validity	.745	.017	.025		.027
	Reliability	.875	.019	.005		.006
A2	Validity	.674	.032	.035	.024	
	Reliability	.883	.014	.011	.014	

*Upper and Lower 27% chosen on the basis of aptitude scores S=R.

TABLE 4

MEANS AND STANDARD DEVIATIONS OF THE CHOICE WEIGHTS*

		Group					
		M1	M2	(M1 + M2)	A1	A2	(A1 + A2)
Correct Responses (n = 80)	Mean	2.3	1.9	1.9	2.5	2.2	2.1
	S.D.	1.4	1.2	1.7	1.4	1.4	1.6
Distracters (n = 240)	Mean	-4.7	-4.4	-4.9	-4.9	-4.2	-4.6
	S.D.	2.6	2.8	2.4	2.6	2.8	2.3
Total (n = 320)	Mean	-3.0	-2.8	-3.2	-3.1	-2.6	-2.9
	S.D.	3.8	3.7	3.7	3.9	3.8	3.6

*Upper and lower 27% based on aptitude test scores.

TABLE 5

DESCRIPTIVE DATA ON SUBJECTS

	Aptitude Test (S = R)			Achievement Test (S = R)		
Group	$\bar{X}$	S.D.	r*	$\bar{X}$	S.D.	r*
M1	53.8	11.8	.89	18.1	6.8	.79
M2	59.0	10.7	.87	22.6	6.9	.82
A1	51.6	12.0	.88	17.3	7.6	.86
A2	54.4	10.9	.88	19.3	7.2	.82

*Odd-even Spearman-Brown estimates of the reliability of the achievement tests.